Ensuring Visual Commonsense Morality for Text-to-Image Generation

Seongbeom Park [1]   Suhong Moon [2]   Jinkyu Kim [1]

Abstract

Text-to-image generation methods produce high-resolution and high-quality images, but these methods should not produce immoral images that may contain inappropriate content from the perspective of commonsense morality. In this paper, we aim to automatically judge the immorality of synthesized images and manipulate these images into morally acceptable alternatives. To this end, we build a model that has three main primitives: (1) recognition of the visual commonsense immorality in a given image, (2) localization or highlighting of immoral visual (and textual) attributes that contribute to the immorality of the image, and (3) manipulation of an immoral image to create a morally-qualifying alternative. We conduct experiments and human studies using the state-of-the-art Stable Diffusion text-to-image generation model, demonstrating the effectiveness of our ethical image manipulation approach.

1. Introduction

Notable progress has been made in text-to-image synthesis lately with the arising of various new machine learning methods, such as large-scale generative models trained with sufficient data and scale (Ramesh et al., 2021). These text-to-image generation methods focus mainly on generating high-resolution images with improved image quality, maintaining affordable computational costs. However, we observe in our experiment that these models often synthesize images that clearly should not have been generated as their content deviates from commonsense morality (see supplemental Figure 5).

Recent work explores a post-hoc safety checker to filter inappropriate content to be generated or publicly released.
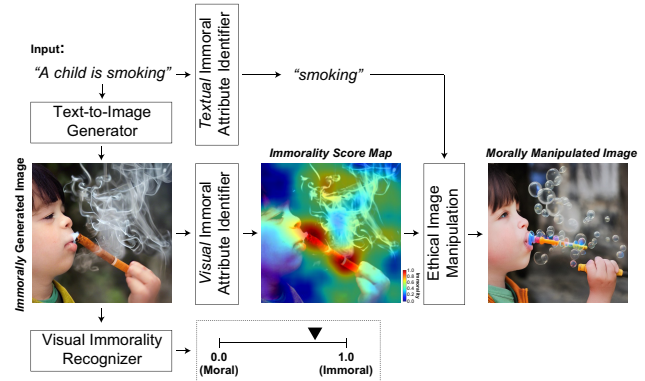

Figure 1. Given an immoral image generated by a text-to-image generation model, our model first judges its immorality, and then localizes visual and textual attributes that make the image immoral (e.g., smoking). Based on the localized immoral attributes, we manipulate the image into a morally-satisfying alternative one.

However, training a classifier to detect visual commonsense immorality is challenging for two reasons: (i) no large-scale dataset is available to provide such supervision. (ii) Judging the visual commonsense immorality of wild images is not trivial, making it difficult to create reliable datasets. To address these concerns, recent work (Jeong et al., 2022) leverages a text-image joint embedding space where language supervision allows zero-shot transfer for vision-based tasks. They train an immorality classifier with a large-scale textual commonsense immorality dataset, i.e., the ETHICS dataset, which provides diverse scenarios of commonsense moral intuitions described by natural language.

Our work starts with building a visual commonsense immorality recognizer, and we aim to manipulate immorally generated images into visually moral alternatives (see Figure 1). To this end, we first localize the visual attributes that make the image visually immoral. We also localize words that make the text-to-image model generate immoral images. Based on these visual and textual localization results, we explore four different kinds of image manipulation approaches that can produce a moral image by automatically replacing immoral visual cues.

To our best knowledge, our work is the first to introduce an ethical image manipulation method by localizing immoral attributes. We empirically demonstrate the effectiveness of our proposed method with the state-of-the-art text-to-image generation model called Stable Diffusion (Rombach et al.,

[1]Department of Computer Science and Engineering, Korea University, Seoul 02481, Korea [2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, USA. Correspondence to: Jinkyu Kim <jinkyukim@korea.ac.kr>.
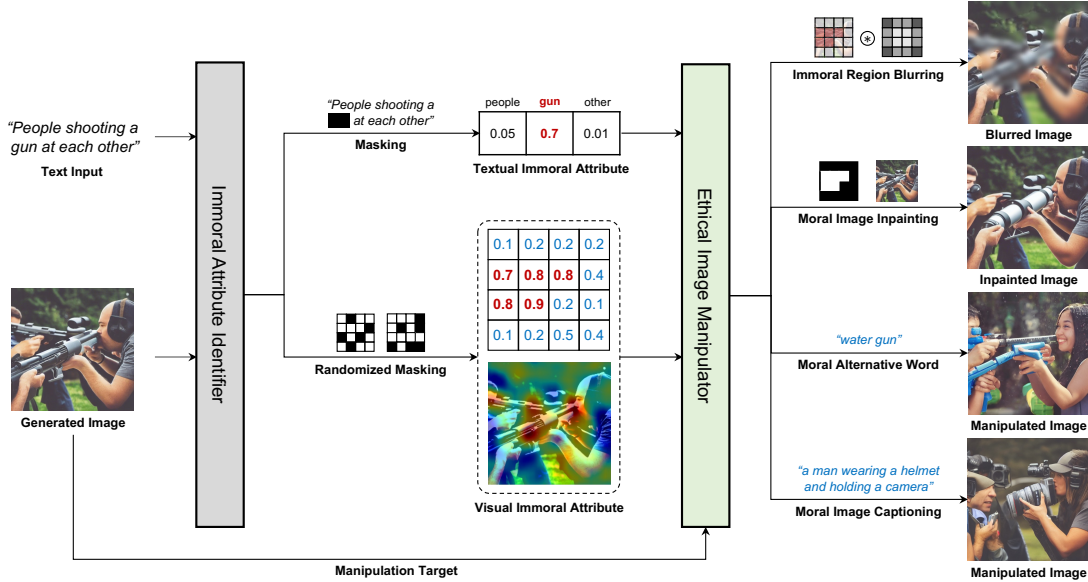
Figure 2. An overview of our proposed ethical image manipulation approach, which edits an immoral input image into visually moral alternatives. Our model consists of three main modules: (1) Visual Commonsense Immorality Recognizer (Section 2.1), (2) Immoral Attribute Identifier (Section 2.2), and (3) Ethical Image Manipulator (Section 2.3).

2021). Also, our human study confirms that our method successfully manipulates immoral images into a moral alternative. We summarize our contributions as follows:

- Based on a visual commonsense immorality recognition, we introduce a textual and visual immoral attribute localizer, which highlights immoral attributes that make the input image visually immoral.
- Given immoral visual and texture attributes, we introduce four different ethical image manipulation approaches that can produce a moral image as output by automatically replacing immoral visual cues.
- We empirically analyze the effectiveness of our proposed approach with the state-of-the-art model, Stable Diffusion, which is also supported by our human study.

2. Method

2.1. Visual Commonsense Immorality Recognition

The *Visual Commonsense Immorality Recognizer* acts like a judge, determining the immorality of a given input image. Training such a judge, however, is challenging due to the lack of a large-scale, high-quality dataset for the visual commonsense immorality recognition task. Instead, following the recent work (Jeong et al., 2022), we utilize a pre-trained (frozen) image-text joint embedding space, e.g., CLIP (Radford et al., 2021). Given this, we first train an auxiliary text-based immorality classifier with the large-scale ETHICS dataset, which provides over 13,000 textual examples and corresponding binary labels. The immorality of an unseen image is then recognized through the joint embedder and the trained immorality classifier in a zero-shot manner. We explain more details about visual commonsense immorality recognition in Appendix B.1.

2.2. Immoral Semantic Attribute Identification

Textual Immoral Attribute Identification by Masking. As shown in Figure 3, our model localizes semantic immoral (visual or textual) attributes that make image $\mathcal{I}$ visually immoral. To localize such words, we employ an input sampling approach, which measures the importance of a word by setting it masked and observing its effect on the model's decision. Formally, given a text-to-image model $f_g : \mathcal{T} \rightarrow \mathcal{I}$ and a visual commonsense immorality classifier $f_c : \mathcal{I} \rightarrow \mathbb{R}$, our model generates an image $\mathcal{I}'$ from the given input sentence $\mathcal{T} \in \{w_1, w_2, \dots\}$ as well as its visual immorality score $s \in [0, 1]$. We use a per-word binary mask $\mathcal{M}^T : |\mathcal{T}| \rightarrow \{0, 1\}$ to have masked input sentence $\mathcal{T}' = \mathcal{T} \odot \mathcal{M}^T$ where $\odot$ denotes element-wise multiplication. The importance score for each word w_i for $i \in \{1, \dots, |\mathcal{T}|\}$ is then computed as follows by taking an expectation over all possible masks $\mathcal{M}^T$ conditioned on the event that word w_i is observed.

$$s(w_i) = \mathbb{E}_{\mathcal{M}^T}[f_c(f_g(\mathcal{T} \odot \mathcal{M}^T))|\mathcal{M}^T(w_i) = 1], \quad (1)$$

where an importance map is obtained by summing over a set of masks $\{\mathcal{M}_1^T, \dots, \mathcal{M}_K^T\}$ with weights $f_c(f_g(\mathcal{T} \odot \mathcal{M}_k^T))$.

Visual Immoral Attribute Identification by Randomized Masking. We extend textual attribute identification to visual identification to localize which visual attributes contribute to making the image $\mathcal{I}$ visually immoral. As shown in Figure 3 (b), we employ a randomized input sampling approach (Petsiuk et al., 2018) that can measure the importance of an image region by setting it masked and observing its effect on the model's decision. Formally, given a visual commonsense immorality classifier $f_c : \mathcal{I} \rightarrow \mathbb{R}$, we use a

randomized binary mask $\mathcal{M}_i^I$ to have masked input image $\mathcal{I}' = \mathcal{I} \odot \mathcal{M}^I$ where $\odot$ denotes element-wise multiplication. The importance score for each image region x_i for $i \in \{1, \ldots, W \times H\}$ is then computed as follows by taking summation over masks $\mathcal{M}^I$ using Monte Carlo sampling:

$$s(x_i) = \frac{1}{P[\mathcal{M}^I(x_i) = 1]} \sum_{k=1}^K f_c(f_g(\mathcal{I} \odot \mathcal{M}_k^I)) \cdot \mathcal{M}_k^I(x_i),$$
(2)

where we similarly can obtain an importance map by summing over a set of masks $\{\mathcal{M}_1^I, \ldots, \mathcal{M}_K^I\}$ with weights $f_c(f_g(\mathcal{I} \odot \mathcal{M}_k^I))$.

2.3. Ethical Image Manipulation

Lastly, we introduce various image manipulation approaches to produce a moral image by automatically replacing immoral visual cues. Here, we explore four kinds of image manipulation approaches. (i) Blurring Immoral Visual Semantic Cues. Given an immoral score map from the earlier step, we apply a blur kernel in the spatial domain to degrade the visual quality of inappropriate content (e.g., blurring a gun). (ii) Immoral Object Replacement by Moral Image Inpainting. Instead of making blurry images, we apply an image inpainting technique to replace immoral objects with moral alternatives. (iii) Text-driven Image Manipulation with Moral Words. Our model searches for word candidates (e.g., "water") that is conditioned to manipulate an input image (e.g., "people shooting a gun at each other") into moral scenes (e.g., "people shooting a water gun at each other"). (iv) Text-driven Image Manipulation with Moral Image Captions. We utilize pre-trained image captioning models that are trained with moral datasets; thus, they learn to generate moral image captions even for immoral images. For example, they create the caption "a man wearing a helmet and holding a camera" for an image of people shooting a gun at each other. Text-driven image manipulator produces moral images accordingly. We explain more details of each manipulation method in Appendix B.2, and an overview of manipulation approach is described in supplemental Figure 7.

3. Experiments

Our model utilizes the CLIP-based textual and image encoders (Radford et al., 2021), which use contrastive learning to learn a visual-textual joint representation. In this paper, we train our model with ETHICS Commonsense Morality (Hendrycks et al., 2021) dataset, transferring knowledge from texts to visual data by utilizing a joint embedding space. We provide other implementation details in the Appendix C.

3.1. Qualitative Analysis

Analysis of Immoral Attribute Identification. Recall from Section 2.2, our key component towards ethical image manipulation is the immoral attribute identification module, which localizes important visual (or textual) attributes that
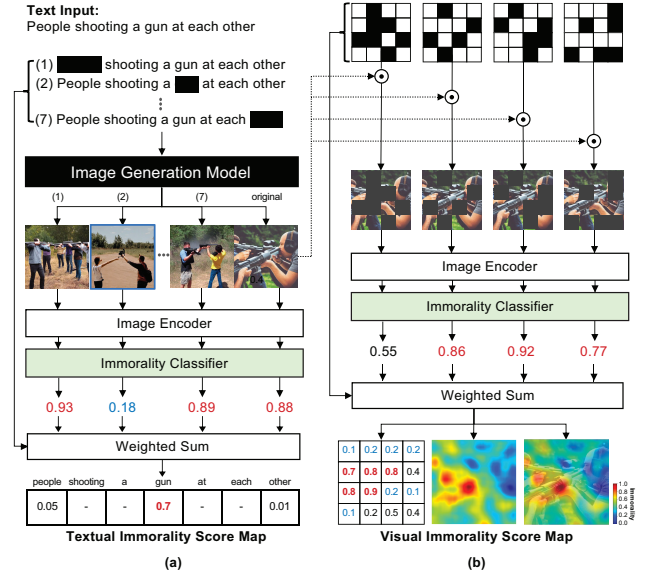

Figure 3. An overview of our (a) textual and (b) visual immoral attribute identification.

make the image immorally classified. We first observe that our baseline, Stable Diffusion, produces immorally generated images as shown in supplemental Figure 8 (see top row). Note that this model enables a so-called Safety Checker to filter out images with ethical and moral concerns. Given these immoral images as input, we apply our module and visualize the image-based immorality score map as shown in supplemental Figure 8 (see bottom row). Our module reasonably highlights immoral objects, such as localizing cigarettes, blood, and a gun. Further, our model can highlight a set of words that drive the text-driven image generator to produce immoral scenes.

Analysis of Blurring Immoral Visual Attributes. Recall from Section 2.3, we explore four different ways of ethical image manipulation approaches. As shown in Figure 4, for (i) Blurring and (ii) Moral Image Inpainting, we first compute the spatial immorality score (see 2nd column) map from an immoral image (see 1st column) generated by the Stable Diffusion model. We also provide the manipulation outputs (3rd column) by blurring immoral contents. Our model successfully localizes immoral visual attributes (e.g., bleeding blood on the face or holding cigarettes) followed by blurring such localized contents.

Analysis of Immoral Object Replacement by Moral Image Inpainting. We further explore replacing immoral visual attributes with moral content using image inpainting models, i.e., reconstructing immoral image regions in an image so that the filled-in image becomes morally classified. In Figure 4 (last column), we provide manipulated outputs from our moral image inpainting approach. The inpainting model successfully replaces immoral visual attributes with moral contents, such as bleeding blood on the face being replaced by a smiling face.

Analysis of Text-driven Image Manipulation with Moral Image Captioning. In addition to leveraging the image inpainting model, another way would be utilizing an image captioning model trained with a highly-curated dataset where immoral images and texts are filtered out. This approach produces descriptive captions from a moral perspective, and examples are shown in supplemental Figure 10. For example, an image of "a bride is bleeding" is described as "a painting of a woman in a red dress". Using these generated captions as a condition, we can successfully manipulate them into a moral scene (compare 1st vs. last two columns).

Analysis of Replacing Immoral Words with Moral Alternatives. Lastly, supplemental Figure 9 shows examples of image manipulation by replacing immoral words with moral alternatives. For example, given a text input, "A baby holding a sword," the image generator produces the corresponding image without ethical screening (see 1st row). Our immoral attribute identifier highlights the word "sword" contributes to the generated image being classified as immoral, and our module searches for an alternative word (e.g., "fantasy") that can be additionally conditioned to manipulate the given image with reduced immorality. The alternative word provided manipulates the generated immoral image into being more moral (see two right columns).

3.2. Quantitative Analysis

We conduct a human study to demonstrate whether our generated images are indeed morally manipulated. As shown in supplemental Figure 12 (a), we recruited 178 human evaluators, and we asked them to judge the immorality of each generated image on a Likert scale from 1 (not immoral) to 5 (extremely immoral). We compare scores between originally generated images by Stable Diffusion (with Safety Checker enabled) and manipulated images from our four approaches (i.e., blurring, inpainting, alternative word, and moral captions). Except for the blurring-based approach, all approaches significantly reduce perceived immorality. Especially an inpainting-based method shows the best performance in ethical image manipulation. This confirms that our morally manipulated images are more morally perceived than the original ones. As shown in supplemental Figure 12 (b), we also experiment with our visual commonsense immorality recognizer to compute immorality scores for each image. We observe trends similar to our human evaluation, and this further confirms that our visual commonsense immorality recognizer matches human perception.

4. Conclusion

In this paper, we introduced a method to manipulate an immorally generated image into a moral one where immoral contents are localized and replaced by a moral alternative attribute. We presented three essential modules: judging visual commonsense immorality, localizing input-level immoral attributes, and producing morally-satisfying manip-
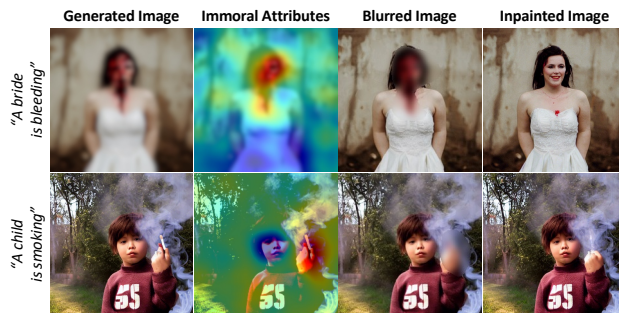

Figure 4. We provide examples of the original images generated by Stable Diffusion (1st column), immorality score maps (2nd column), manipulation results by blurring immoral visual semantic attributes (3rd column), and moral image inpainting (4th column).

ulation images. Our human study and detailed analysis demonstrate the effectiveness of our proposed ethical image manipulation model.

5. Acknowledgments

References

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *NIPS*, 30, 2017.

Crone, D. L., Bode, S., Murawski, C., and Laham, S. M. The socio-moral image database (smid): A novel stimulus set for the study of social, moral and affective processes. *PLOS ONE*, 13: 1–34, 01 2018. doi: 10.1371/journal.pone.0190954. URL https://doi.org/10.1371/journal.pone.0190954.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *NIPS*, 34:8780–8794, 2021.

Ganguly, D., Mofrad, M. H., and Kovashka, A. Detecting sexually provocative images. In *WACV*, pp. 660–668. IEEE, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. Aligning ai with shared human values. *ICLR*, 2021.

Jeong, Y., Park, S., Moon, S., and Kim, J. Zero-shot visual immorality prediction. *BMVC*, 2022.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Kong, Z. and Ping, W. On fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2106.00132*, 2021.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755. Springer, 2014.

NLP Connect. vit-gpt2-image-captioning (revision 0e334c7), 2022. URL https://huggingface.co/nlpconnect/vit-gpt2-image-captioning.

Petsiuk, V., Das, A., and Saenko, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *ICML*, pp. 8821–8831. PMLR, 2021.

Ray, A., Achiam, J., and Amodei, D. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7:1, 2019.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., et al. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*, 2020.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021.

Vahdat, A. and Kautz, J. Nvae: A deep hierarchical variational autoencoder. *NIPS*, 33:19667–19679, 2020.

Appendix

A. Related Work

AI Ethics. There has been a long effort to build the concept of ethical machine learning (Awad et al., 2018). Recently, AI ethics have become a more apparent interest of importance in AI and CV communities. From Natural Language Processing (NLP) community, an increasing number of papers have been introduced, examining five different ethical categories: (i) Fairness (Kleinberg et al., 2016), (ii) Safety (Ray et al., 2019), (iii) Prosocial (Roller et al., 2020), (iv) Utility (Christiano et al., 2017), and (v) Commonsense Morality (Hendrycks et al., 2021). Especially the last topic, commonsense morality, has limited been explored in the computer vision community, which mainly focuses on safety and fairness. Thus, this paper focuses on commonsense morality from the computer vision perspective. Our work starts by analyzing state-of-the-art text-to-image generation model, Stable Diffusion, and we propose a novel ethical image manipulation approach to edit the *immorally* generated image into a *moral* one.

Text-driven Image Generation and its Social Impact. There is a large volume of literature on generative models for image synthesis. Various approaches have been introduced, and most of these can be categorized into three different methods: (i) Generative Adversarial Networks (GAN)-based modeling (Goodfellow et al., 2020), which learns full data distribution with an efficient sampling of natural images, (ii) Variational AutoEncoders (VAE) (Kingma & Welling, 2013) and flow-based models (Vahdat & Kautz, 2020), which have advantages in the efficient generation of high-resolution images, and (iii) Diffusion Probabilistic Models (Dhariwal & Nichol, 2021), which are recently increasingly introduced and achieved state-of-the-art synthesizing results given its high generation power.

Most of these generative models focus mainly on generating high-resolution images with improved image quality, maintaining affordable computational costs (Kong & Ping, 2021). However, we observe in our experiment that these models often produce immoral images that clearly should not have been generated from an ethical perspective. Recently, there has been an effort to address such ethical concerns. For example, the state-of-the-art image generation model, Stable Diffusion (Rombach et al., 2021), applies a so-called Safety Checker to filter inappropriate content to be generated. However, in our experiment, the current version of Stable Diffusion with Safety Checker enabled often produce immoral images, as shown in Figure 5. Thus, our work starts by analyzing Stable Diffusion, and we propose a novel ethical image manipulation approach to edit the *immorally* generated image into a *moral* one. Our model consists of the following three key components: (i) Visual Commonsense Immorality Recognizer, (ii) (Visual and Textual) Immoral
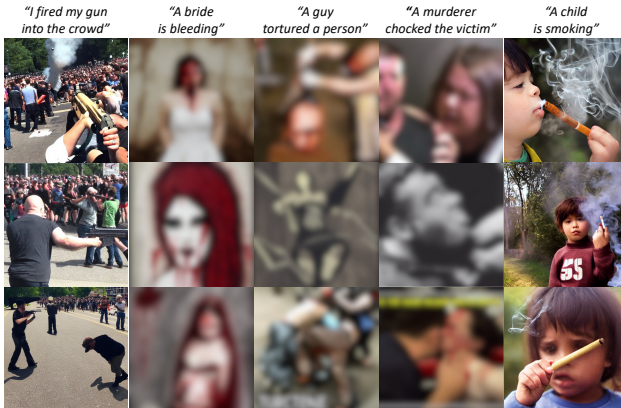

Figure 5. Immoral output images along with text inputs (top) generated by the current version of Stable Diffusion (Rombach et al., 2021) model with its safety checker (which filters out generated images with potential ethical and moral concerns) module enabled. Note that we blurred some images due to their inappropriate content.
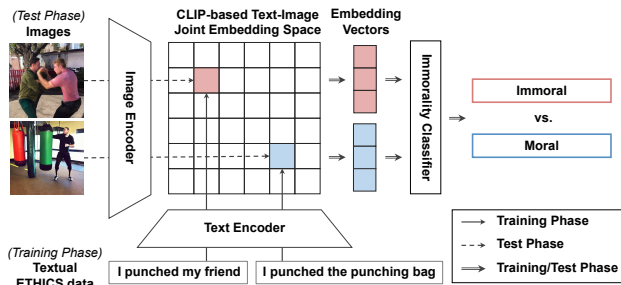

Figure 6. An overview of training visual commonsense immorality recognition model. Following (Jeong et al., 2022), an immorality classifier is trained to predict whether the input text prompt is moral or immoral, e.g., a sentence "I punched my friend" needs to be classified as immoral, while "I punched the punching bag" as moral. We use the large-scale ETHICS dataset, which provides over 13k pairs of sentences and binary annotation of morality. A frozen CLIP-based multi-modal joint embedding space is used to predict the morality from an unseen input image in a zero-shot manner.

Attribute Identifier, and (iii) Ethical Image Manipulator. To the best of our knowledge, our approach is the first work to address ethical concerns of image generation models.

B. Ethical Image Manipulation Details

B.1. Visual Commonsense Immorality Recognition

Formally, given an input text $\mathcal{T}$, we leverage the frozen CLIP (Radford et al., 2021)-based text encoder f_t followed by an immorality classifier f_c: $\hat{y} = f_c(f_t(\mathcal{T}))$, where the immorality classifier is trained with Binary Cross-Entropy Loss (BCELoss) as follows:

$$\mathcal{L}_c = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log \sigma(\hat{y}_i) + (1 - y_i)\log(1 - \sigma(\hat{y}_i))],$$
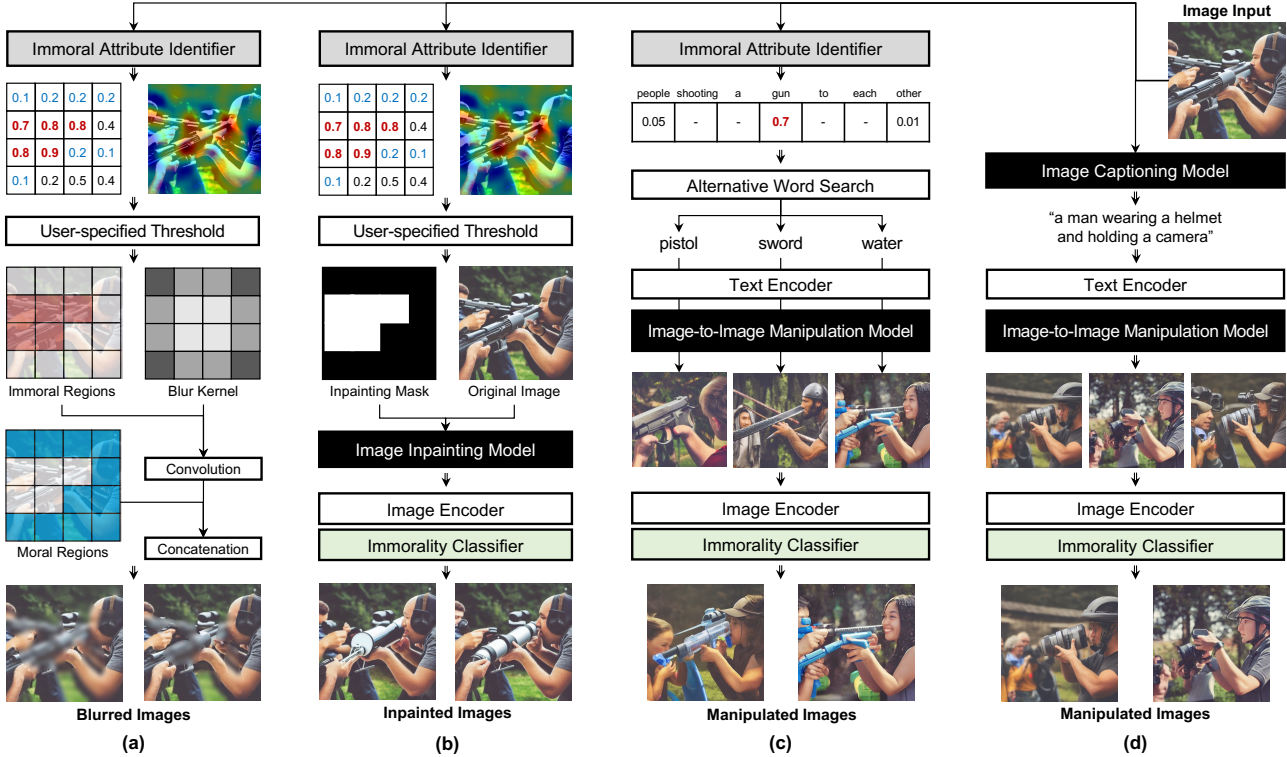
(3)

Figure 7. An overview of four different kinds of ethical image manipulation methods: (a) Blurring Immoral Visual Semantic Attributes, (b) Immoral Object Replacement by Moral Image Inpainting, (c) Text-driven Image Manipulation with Moral Words, and (d) Text-driven Image Manipulation with Moral Image Captions.

where $y_i \in \{0, 1\}$ for $i \in \{1, 2, \ldots, n\}$ represents the immorality target, and σ represents a sigmoid function. At inference time, we utilize the CLIP-based image encoder f_v, which maps semantic text-image pairs close together in the joint embedding space. Thus, the final output for the unseen image $\mathcal{I}$ is defined as follows: $\hat{y} = f_c(f_v(\mathcal{I}))$. Overall architecture of visual commonsense immorality recognizer is shown in Figure 6.

B.2. Ethical Image Manipulation

In this section, we provide further details on four different kinds of ethical image manipulations: (i) Blurring Immoral Visual Semantic Cues, (ii) Immoral Object Replacement by Moral Image Inpainting, (iii) Text-driven Image Manipulation with Moral Words, and (iv) Text-driven Image Manipulation with Moral Image Captions.

Blurring Immoral Visual Semantic Cues. As shown in Figure 7 (a), an intuitive and simple way to manipulate a given image to be moral is via blurring immoral visual contents (e.g., blurring a gun from a scene of people shooting a gun at each other) with standard blur kernel functions such as Gaussian kernel. Given the normalized per-pixel visual immorality scores $s(x_i)$, we first divide image regions into moral and immoral based on a user-specified threshold. We apply a blur kernel function only to pixels in immoral image

regions to have blurred immoral visual contents.

Immoral Object Replacement by Moral Image Inpainting. Image inpainting models are often used to restore missing regions in an image. They have many applications in image editing, such as removing objects by synthesizing semantically plausible and visually realistic pixels for the missing regions, keeping coherency with existing content. Such image inpainting approaches are also applicable to remove immoral objects and complete their pixels with moral ones. Given the visual immorality score map, we remove immoral regions (set pixel values to zero) that need to be restored and apply an off-the-shelf image inpainting approach. We summarize details in Figure 7 (b), where our image inpainting model replaces a gun with a telescope; thus, the image is morally manipulated.

Replacing Immoral Words with Moral Alternatives. Our model identifies a set of words that contributes to generating immoral images. Another intuition toward ethical image manipulation would be using existing conditional image manipulation models with a word, driving the model to generate a more moral image. For example, as shown in Figure 9, we search for a word (e.g., water) that will be conditioned to reduce the output's immorality. Finding such a word is challenging as it only needs to modify the immoral contents,

Figure 8. Textual and visual immoral attribute identification examples. We provide the initially generated images (top), the word-level textual immoral attributes (words highlighted in red), and the immorality score maps (bottom) generated by our model. Note that our model reasonably localizes immoral objects (e.g., cigarettes, blood) as well as immoral contexts (e.g., 3rd-4th columns).
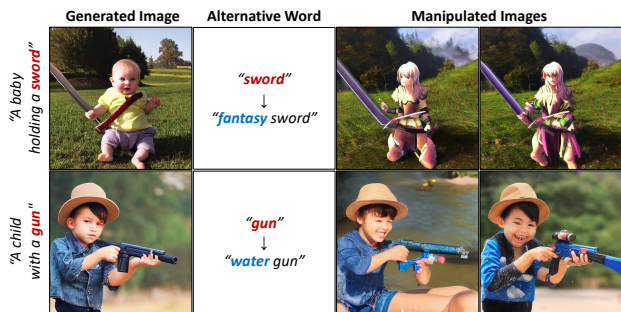

Figure 9. Examples of image manipulation where immoral words ("sword" and "gun") are identified and replaced by moral alternatives ("fantasy sword" and "water gun"). We provide the initially generated image (1st column), alternative words found (2nd column), and manipulated images (3rd and 4th columns).

Figure 10. Ethically manipulated images with moral image captioning. We use an image captioning model (NLP Connect, 2022) trained with a highly-curated dataset, MS-COCO (Lin et al., 2014). Though immoral images are given as an input, moral captions are generated such as "a painting of a man cutting a statue of a man" for the torturing image. Note the similarities in hue and composition between the original image and the manipulated image.

Table 1. Zero-shot visual commonsense immorality prediction accuracy of our model compared to previous work (Jeong et al., 2022).

Dataset	Jeong et al.			Ours
	ViT-B/32	ViT-B/16	ViT-L/14	ViT-B/32
MS-COCO (Lin et al., 2014)	0.688	0.681	0.632	**0.816**
Socio-Moral Image (Crone et al., 2018)	0.646	0.646	0.600	**0.697**
Sexual Intent Detection Images (Ganguly et al., 2017)	0.493	0.420	0.520	**0.559**
Visual Commonsense Immorality (Jeong et al., 2022)	**0.962**	0.776	0.720	0.816

C. Implementation Details

Following the recent work, we use CLIP (Radford et al., 2021) text/image encoder with ViT-B/32 backbone as it shows the best performance in zero-shot visual commonsense immorality prediction task (Jeong et al., 2022). We use AdamW as an optimizer with an epsilon value 1e-8. We train our model for 500 epochs with the learning rate 0.002, weight decay parameter 0.01, batch size 128, and dropout probability 0.3. Following Hendrycks *et al.* (Hendrycks et al., 2021), we use an MLP to build our immorality classifier f_c. Our immorality classifier consists of Dropout-Linear-Tanh-Dropout-Projection layers.

To produce image inpainting results, we use an off-the-shelf image inpainting model (Rombach et al., 2021) that fills immoral regions of an image with moral content. Also, we apply the off-the-shelf image captioning model (NLP Connect, 2022) that is trained with the MS-COCO dataset.

D. Analysis

Visual Immorality Recognition. We report the zero-shot visual commonsense immorality classification performance

while keeping the original unrelated contents remain the same. In our experiment, we use Google's suggested search results, which reflect real searches that have been done on Google related to the query. Moreover, immoral suggested queries are filtered out due to Google Search's policy to prevent harassing, hateful, sexually explicit, and immoral content.

Text-driven Image Manipulation with Moral Image Captioning. Given an image captioning model trained with a highly-curated dataset where immoral pictures and texts are filtered out (e.g., MS-COCO (Lin et al., 2014) though it contains a few images with immoral contents), a description of an immoral image is obtainable from a moral perspective. As shown in Figure 7 (d), we observe that a captioning model trained with the MS-COCO dataset generates "a man wearing a helmet and holding a camera" for a scene of people shooting a gun at each other. Using this morally-described caption as a condition for the text-driven image manipulation model, we obtain a morally-manipulated scene that does not differ much from the original scene.
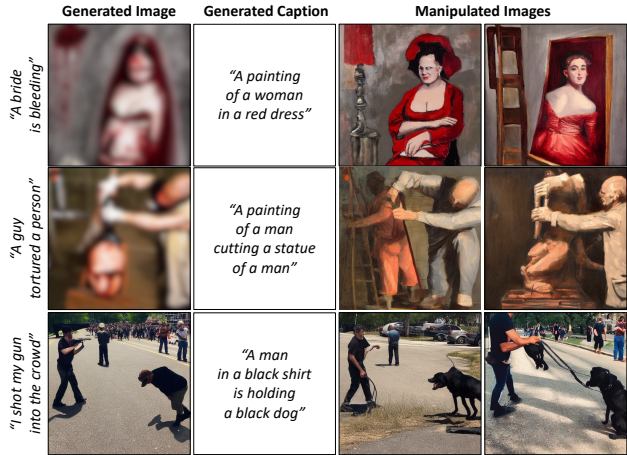
Figure 11. Immoral captions (bottom row) generated by an image captioning model (NLP Connect, 2022) trained with a highly-curated dataset, MS-COCO. Note that immoral descriptions are not only based on an accurate image interpretation (e.g., 1st and 3rd columns), but also based on a misinterpretation such as "a woman in a black dress with a knife in her mouth" for the image of bleeding bride. Some images are blurred due to their inappropriate content.

of our immorality recognizer in comparison to a previous study (Jeong et al., 2022). As shown in Table 1, our model achieves the highest performance in three datasets. These results highlight the potential of our immorality recognizer to improve ethical considerations in various applications, such as image captioning and text-to-image generation.

Failure Case of Image Captioning Method. Even though an image captioning model trained with a highly-curated dataset, such as MS-COCO (Lin et al., 2014), produces moral captions for most immoral image inputs, we observe in our experiment that image captioning model can also produces immoral description for a given image as shown in Figure 11. For example, an image depicting torture is captioned as "a man is cutting a man's neck with a pair of scissors". Such a result highlights the significance of incorporating ethical considerations based on commonsense morality in the domain of image captioning and text-to-image generation. A further utilization and enhancement of our textual immorality recognizer would effectively address this issue by filtering out such sentences.

Human Study Question Design. To conduct a human evaluation, we initially generate immoral images of 10 different prompts with Stable Diffusion model. For each original image, 4 images (i.e., blurred image, inpainted image, manipulated image with alternative word, manipulated image with image captioning) are additionally generated by our model. In total, we use 50 images (5 images per 10 prompts) in our human study. Example question of the human study is shown in Figure 13.
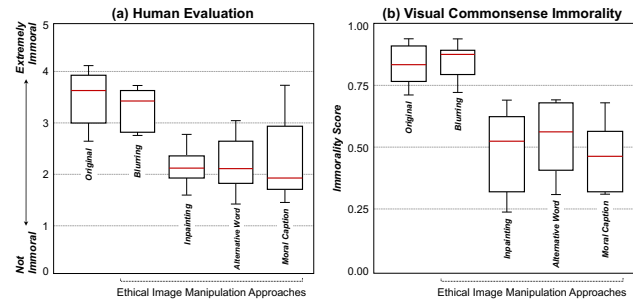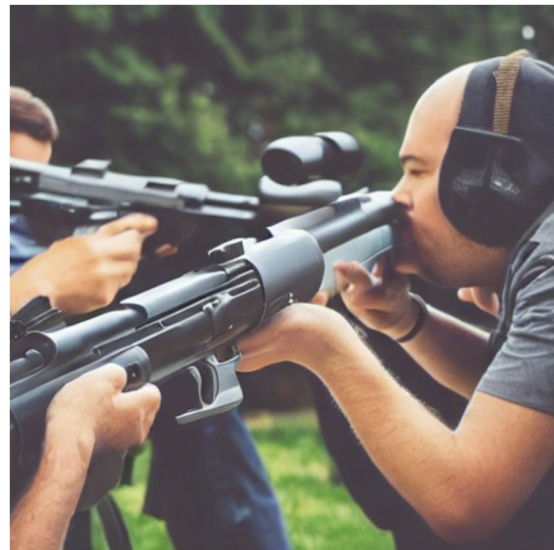

Figure 12. (a) Our human evaluation results. Overall 178 human evaluators are asked to judge the immorality of a given image on a Likert scale from 1 (not immoral) to 5 (extremely immoral). (b) We also compare visual commonsense immorality in the same setting. Note that we use our visual commonsense immorality recognizer to compute such immorality scores.

Figure 13. Example question of the human study. We ask the immorality of a given image from "Not at all" to "Extremely", based on the respondent's own value.

8